13 / PRTS

Method and Devices for the Co-articulation-specific Concatenation of Audio Segments

The invention relates to a method and a device for the concatenation of audio segments for the generation of synthesised acoustical data, in particular synthesised speech. In particular, the invention relates to synthesised voice signals which have been generated by the inventive co-articulation-specific concatenation of voice segments, as well as to a data carrier which contains a computer program for the inventive generation of synthesised acoustical data, in particular, synthesised speech.

In addition, the invention relates to a data storage which contains audio segments which are suited for the inventive co-articulation-specific concatenation, and a sound carrier which, according to the invention, contains synthesised acoustical data.

It must be emphasised that both the state of the art represented in the following, and the present invention relate to the entire field of the synthesis of acoustical data by means of the concatenation of individual audio segments which are obtained in any manner. However, for the sake of simplifying the discussion of the state of the art as well as the description of the present invention, the following explanations refer specifically to synthesised voice data by means of the concatenation of individual voice segments.

During the past years, the data-based approach has been successful over the rule-based approach in the field of speech synthesis, and can be found in various methods and systems for speech synthesis. Although the rule-based approach principally enables a better speech synthesis, it is necessary for its implementation to explicitly phrase the entire knowledge which is required for speech generation, i.e. to formally model the speech to be synthesised. Due to the fact that the known speech models comprise a simplification of the speech to be synthesised, the voice quality of the speech generated in this manner is not sufficient.

For this reason, a data-based speech synthesis is carried out to an increasing extent, wherein corresponding segments are selected from a database containing individual voice segments and linked (concatenated) to each other. In this context, the voice quality is primarily depending on the number and type of the available voice segments, because only that speech can be synthesised which is reproduced by voice segments in the database. In order to minimise the number of the voice segments to be provided and, nevertheless, to still generate a high quality synthesised speech, various methods are known which carry out a linking (concatenation) of the voice segments according to complex rules.

When using such methods or corresponding devices, respectively, an inventory, i.e. a database comprising the voice audio segments can be employed which is complete and manageable. An inventory is complete if it is capable of generating any sound sequence of the speech to be synthesised, and it is manageable if the number and type of the data of the inventory can be processed in a desired manner by means of the technically available means. Furthermore, such a method must ensure that the concatenation of the individual inventory elements generates a synthesised speech which differs as little as possible from a naturally spoken speech. To this end, a synthesised speech must be fluent and comprise the same articulatory effects as a natural speech. In this context, the so-called co-articulatory effects, i.e. the mutual influence of phones, are of particular importance. For this reason, the inventory elements should be of such a nature that they consider the co-articulation of individual successive phones. In addition, a method for the concatenation of the inventory elements should link the elements, even beyond word and phrase boundaries, under consideration of the co-articulation of individual successive phones as well as of the higher-order co-articulation of several successive phones.

Before presenting the state of the art, a few terms from the field of speech synthesis, which are necessary for a better understanding, will be explained in the following:

- A phone is a class of any sound events (noises, sounds, tones, etc.). The sound events are classified in accordance with a classification scheme into phone classes. A sound event belongs to a phoneme if the values of the sound event are within the range of values defined for the phone with respect to the parameters (e.g. spectrum, tone level, volume, chest or head voice, co-articulation, resonance cavities, emotion, etc.) used for the classification.
The classification scheme for phones depends on the type of application. For vocal sounds (= phones), the IPA classification is generally used. However, the definition of the term phone as used herein is not limited to this, but any other parameters can be used. If, for example, in addition to the IPA classification, the tone level or the emotional expression are included as parameters in the classification, two 'a' phones with different tone level or different emotional expression become different phones in the sense of the definition. Phones can, however, also be the tones of a musical instrument, e.g. a violin, in the different tone levels and the different modes of playing (up-bow and down-bow, detaché, spiccato, marcato, pizzicato, col legno, etc.). Phones can be the barking of dogs or the squealing of a car door.

Phones can be reproduced by audio segments which contain corresponding acoustical data.

In the description of the invention following the definitions, the term vocal sound can invariably be replaced by the term phone in the sense of the previous definition, and the term phoneme can be replaced by the term phonetic character. (This also applies the other way round, because phones are vocal sounds classified according to the IPA classification).

- A static phone has bands which are similar to previous or subsequent bands of the static phone. The similarity need not necessarily be an exact correspondence as in the periods of a sinusoidal tone, but is analogous to the similarity as it prevails between the bands of the static phones defined in the following.

- A dynamic phone has no bands with a similarity with previous or subsequent bands of the dynamic phone, such as, e.g. the sound event of an explosion or a dynamic phone.

- A phone is a vocal sound which is generated by the organs of speech (a vocal sound). The phones are classified into static and dynamic phones.

- The static phones include vowels, diphtongs, nasals, laterals, vibrants, and fricatives.

- The dynamic phones include plosives, affricates, glottal stops, and click sounds.

- A phoneme is the formal description of a phone, with the formal description usually being effected by phonetic characters.

- The co-articulation refers to the phenomenon that a sound, i.e. a phone, too, is influenced by upstream or downstream sounds or phones, respectively, with the co-articulation occurring both between immediately neighbouring sounds/phones, but also covering a sequence of several sounds/phones as well (for example in rounding the lips).

A sound or phone, respectively, can therefore be classified into three bands (see also Fig. 1b):

- The initial co-articulation band comprises the band from the start of a sound/phone to the end of the co-articulation due to a upstream sound/phone.

- The solo articulation band is the band of the sound/phone which is not influenced by an upstream or downstream sound or an upstream or downstream phone, respectively.

- The end co-articulation band comprises the band from the start of the co-articulation due to a downstream sound/phone to the end of the sound/phone.

- The co-articulation band comprises an end co-articulation band and the neighbouring initial co-articulation band of the neighbouring sound/phone.

- A polyphone is a sequence of phones.

- The elements of an inventory are audio segments stored in a coded form which reproduce sounds, portions of sounds, sequences of sounds, or portions of sequences of sounds, or phones, portions of phones, polyphones, or portions of polyphones, respectively. For a better understanding of the potential structure of an audio segment/inventory element, reference is made to Fig. 2a which shows a conventional audio segment, and Figs. 2b - 2l which show inventive audio segments. In addition, it should be mentioned that audio segments can can be formed from smaller or larger audio segments which are included in the inventory or a database. Furthermore, audio segments can also be provided in a transformed form (e.g. in a Fourier-transformed form) in the inventory or the database. Audio segments for the present invention can also come from a prior synthesis step (which is not part of the method). Audio segments include at least a part of an initial co-articulation band, a solo articulation band, and/or an end co-articulation band. In lieu of audio segments, it is also possible to use bands of audio segments.

- The term <u>concatenation</u> implies the joining of two audio segments.

- The <u>concatenation instance</u> if the point of time in which two audio segments are joined.

The concatenation can be effected in various ways, e.g. with a <u>cross fade</u> or a <u>hard fade</u> (see also Figs. 3a - 3e):

- In a <u>cross fade</u>, a downstream band of a first audio segment band and an upstream band of a second audio segment band are processed by means of suitable transfer functions, and subsequently these two bands are overlappingly added in such a manner that at the most the shorter band with respect to time of the two bands is completely overlapped by the longer one with respect to time of the two band.

- In a <u>hard fade</u>, a later band of a first audio segment and an earlier band of a second audio segment are processed by means of suitable transfer functions, with the two audio segments being joined to one another in such a manner that the later band of the first audio segment and the earlier band of the second audio segment do not overlap.

The co-articulation band is primarily noticeable in that a concatenation therein is associated with discontinuities (e.g. spectral skips).

In addition, reference is to be made that, strictly speaking, a hard fade is a boundary case of a cross fade, in which an overlap of a later band of a first audio segment and an earlier band of a second audio segment has a length of zero. This allows to replace a cross fade with a hard fade in certain, e.g. extremely time-critical applications, with such an approach to be contemplated scrupulously, because it results in considerable quality losses in the concatenation of audio segments which actually are to be concatenated by a cross fade.

- The term prosody refers to changes in the voice frequency and the voice rhythm which occur in spoken words or phrases, respectively. The consideration of such prosodic information is necessary in the speech synthesis in order to generate a natural word or phrase melody, respectively.

From WO 95/30193 a method and a device are known for the conversion of text to audible voice signals under utilising a neural network. For this purpose, the text to be converted to speech is converted to a sequence of phonema by means of a converter unit, with information on the syntactic boundaries of the text and the stress of the individual components of the text being additionally generated. This information, together with the phonema, are transferred to a device which determines the duration of the pronunciation of the individual phonema in a rule-based manner. A processor generates a suitable input for the neural network from each individual phoneme in connection with the corresponding syntactic and time-related information, with said input for the neural network also comprising the corresponding prosodic information for the entire phoneme sequence. From the available audio segments the neural network then selects only those segments which best reproduce the input phonema and links said audio segments accordingly. In this linking operation the individual audio segments with respect to their duration, total amplitude, and frequency are matched to upstream and downstream audio segments under consideration of the prosodic information of the speech to be synthesised and time successively connected with each other. A modification of individual bands of the audio segments is not described therein.

For the generation of the audio segments which are required for this method, the neural network has first to be trained by dividing naturally spoken speech into phones or phone sequences and assigning these phones or phone sequences corresponding phonema or phoneme sequences in the form of audio segments. Due to the fact that this method provides for a modification of individual audio segments only, but not for a modification of individual bands of an audio segment, the neural network must be trained with as many different phones or phone sequences as possible for converting any text to a synthesised speech with a natural sound. Depending of the application, this may prove to require very high expenditures. On the other hand, an insufficient training process of the neural network may have a negative influence on the quality of the speech to be synthesised. Moreover, it is not possible with the method described therein to determine the concatenation instance of the individual audio segments depending on upstream or downstream audio segments, in order to perform a co-articulation-specific concatenation.

US-5,524,172 describes a device for the generation of synthesised speech, which utilises the so-called diphone method. Here, a text which is to be converted to synthesised speech is divided into phoneme sequences, with corresponding prosodic information being assigned to each phoneme sequence. From a database which contains audio segments in the form of diphones, for each phoneme of the sequence two diphones reproducing the phoneme are selected and concatenated under consideration of the corresponding prosodic information. In the concatenation the two diphones each are weighted by means of a suitable filter, and the duration and tone level of both diphones modified in such a manner that upon the linking of the diphones a synthesised phone sequence is generated, whose duration and tone level correspond to the duration and tone level of the desired phoneme sequence. In the concatenation the individual diphones are added in such a manner that a later band of a first diphone and an earlier band of a second diphone overlap, with the instance of concatenation being generally in the area of stationary bands of the individual diphones (see Fig. 2a). Due to the fact that a variation of the instance of concatenation under consideration of the coarticulation of successive audio segments (diphones) is not intended, the quality (naturalness and audibility) of a speech synthesised in such a manner can be negatively influenced.

A further development of the previously discussed method can be found in EP-0,813,184 A1. In this case, too, a text to be converted to synthesised speech is divided into individual phonema or phoneme sequences, and corresponding audio segments are selected from a database and concatenated. In order to achieve an improvement of the synthesised speech, two approaches have been realised with this method, which differ from the state of the art discussed so far. With the use of a smoothing filter which accounts for the lower-frequency harmonic frequency components of an upstream and a downstream audio segment, the transition from the upstream audio segment to the downstream audio segment is to be optimised, in that a later band of the upstream audio segment and an earlier band of the downstream audio segment in the frequency range are tuned to each other. In addition, the database provides audio segments which are slightly different from one another but are suited for synthesising one and the same phoneme. In this manner, the natural variation of the speech is to be mimicked in order to achieve a higher quality of the synthesised speech. Both the use of the smoothing filter and the selection from a plurality of various audio segments for the realisation of a phoneme require a high computing power of the used system components in the implementation of this method. Moreover, the volume of the database increases due to the increased number of the provided audio segments. Furthermore, this method, too, does not provide for a ca-articulation dependent choice of the concatenation instance of individual audio segments, which may reduce the quality of the synthesised speech.

DE 693 18 209 T2 deals with formant synthesis. According to this document two multi-voice phones are connected with each other using an interpolation mechanism which is applied to a last phoneme of an upstream phone and to a first phoneme of a downstream phone, with the two phonema of the two phones being identical and with the connected phones are superposed to one phoneme. Upon the superposition, each of the curves describing the two phonema is weighted with a weighting function. The weighting function is applied to a band of each phoneme, which begins immediately after the start of the phoneme and ends immediately before the end of the phoneme. Thus, in the concatenation of phones described therein, the bands of the phonema, which form the transition between phones, correspond essentially to the respective entire phonema. This means, that portions of the phonema used for concatenation, invariably comprise all three bands, i.e. the respective initial co-articulation band, solo articulation band, and end co-articulation band. Consequently, D1 teaches an approach how the transitions between two phones are to be smoothed.

Moreover, according to this document the instance of the concatenation of two phones is established in such a manner that the last phoneme in the upstream phone and the first phoneme in the downstream phone completely overlap.

Principally, it is to be stated that DE 689 15 353 T2 aims at improving the tone quality, in that an approach is specified how to design the transition between two neighbouring sampling values. This is of particular relevance in the case of low sampling rates.

In the speech synthesis described in this document, waveforms are used which reproduce the phones to be concatenated. With waveforms for upstream phones, a corresponding final sampling value and an associated zero crossing point are established, while with waveforms for downstream phones, a corresponding first upper sampling value and an associated zero crossing point are established. Depending on these established sampling values and the associated zero crossing points, phones are connected with each other by means of maximal four different ways. The number of connection types is reduced to two, if the waveforms are generated by utilising the Nyquist theoreme. DE 689 15 353 T2 describes that the used band of waveforms extends between the last sampling value of the upstream waveform and the first sampling value of the downstream waveform. A variation of the duration of the used bands as a function of the waveforms to be concatenated, as it is the case with the invention, is not disclosed in D1.

In summary, it can be said that the state of the art allows to synthesise any phoneme sequences, but that the phoneme sequences synthesised in this manner do not possess an authentic voice quality. A synthesised phoneme sequence has an authentic voice quality if it cannot be distinguished by a listener from the same phoneme sequence spoken by a real speaker.

Methods are also known which use an inventory which comprises complete words and/or phrases in authentic voice quality as inventory elements. For the speech synthesis, these elements are brought into a desired order, with the possibilities of various voice sequences being limited to a high degree by the volume of such an inventory. The synthesis of any phoneme sequences is not possible with these methods.

It is therefore the object of the present invention to provide a method and a corresponding device which eliminate the problems of the state of the art and enable the generation of synthesised acoustical data, in particular, synthesised voice data, which a listener cannot distinguish from corresponding natural acoustical data, in particular, naturally spoken speech. The acoustical data synthesised by means of the invention, in particular, synthesised voice data, is to possess an authentic acoustical quality, in particular, an authentic voice quality.

For the solution of this object the invention provides a method according to Claim 1, a device according to Claim 14, synthesised voice signals according to Claim 28, a data carrier according to Claim 39, a data storage according to Claim 51, as well as a sound carrier according to Claim 60. The invention therefore makes it possible to generate synthesised acoustical data which reproduces a sequence of phones, in that in the concatenation of audio segments, the instance of the concatenation of two audio segments is determined, depending on properties of the audio segments to be linked, in particular the co-articulation effects which relate to the two audio segments. According to the present invention, the instance of concatenation is preferably selected in the vicinity of the boundaries of the solo articulation band. In this manner, a voice quality is achieved, which cannot be obtained with the state of the art. The required computation power is not higher than with the state of the art.

In order to mimic the variations which can be found in the corresponding natural acoustical data, in the synthesis of acoustical data, the invention provides for a different selection of the audio segment bands as well as for different ways of the co-articulation-specific concatenation. A higher degree of naturalness of the synthesised acoustical data is achieved if a later audio segment band, whose start reproduces a static phone, is connected with an earlier audio segment band by means of a cross fade, or if a later audio segment band, whose start reproduces a dynamic phone, is connected with an earlier audio segment band by means of a hard fade, respectively. In addition, it is advantageous to generate the start of the synthesised acoustical data to be generated by using an audio segment band which reproduces the start of a phone sequence, or to generate the end of the synthesised acoustical data to be generated by using an audio segment band which reproduces the end of a phone sequence, respectively.

In order to carry out the generation of the synthesised acoustical data in a simpler and faster way, the invention makes it possible to reduce the number of audio segment bands which are required for data synthesising, in that audio segment bands are used which always start with the reproduction of a dynamic phone, which allows to carry out all concatenations of these audio segment bands by means of a hard fade. For this purpose, later audio segment bands are connected with earlier audio segment bands whose starts always reproduce a dynamic phone. In this manner, high-quality synthesised acoustical data according to the invention can be generated with low computing power (e.g. in the case of answering machines or car navigation systems).

In addition, the invention provides for mimicking acoustical phenomena which result because of a mutual influence of individual segments of corresponding natural acoustical data. In particular, it is intended here to process individual audio segments or individual bands of the audio segments, respectively, with the aid of suitable functions. Thus it is possible to modify i.a. the frequency, the duration, the amplitude, or the spectrum of the audio segments. If synthesised voice data is generated by means of the invention, then preferably prosodic information and/or higher-order co-articulation effects are taken into consideration for the solution of this object.

The signal characteristic of synthesised acoustical data can additionally be improved if the concatenation instance is set in places of the individual audio segment bands to be connected, where the two used bands are in agreement with each other with respect to one or several suitable properties. These properties can be i.a.: zero point, amplitude value, gradient, derivative of any degree, spectrum, tone level, amplitude value in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

The invention further enables to improve the selection of audio segment bands for the generation of the synthesised acoustical data, as well as to make their concatenation more efficient, in that heuristic knowledge is used which relates to the selection, processing, variation, and concatenation of the audio segment bands.

In order to generate synthesised acoustical data which is voice data which does not differ from corresponding natural voice data, preferably audio segment bands are used which reproduce sounds/phones or portions of sound sequences/phone sequences.

Furthermore, the invention permits the utilisation of the generated synthesised acoustical data, in that this data is convertible to acoustical signals and/or voice signals, and/or storable in a data carrier.

In addition, the invention can be used for providing synthesised voice signals which differ from known synthesised voice signals in that, concerning their naturalness and audibility, they do not differ from real speech. For this purpose, audio segment bands are concatenated in a co-articulation-specific manner, each of which reproduces portions of the sound sequence/phone sequence of the speech to be synthesised, in that the bands of the audio segments to be used as well as the instance of the concatenation of these band are established according to the invention as defined in Claim 28.

A further improvement of the synthesised speech can be achieved if a later audio segment band whose start reproduces a static phone is connected with an earlier audio segment band by means of a cross fade, or if a later audio segment band whose start reproduces a dynamic phone, respectively, is connected with an earlier audio segment band by means of a hard fade. Herein, static phones comprise vowels, diphtongs, liquids, fricatives, vibrants, and nasals, and dynamic phones comprise plosives, affricates, glottal stops, and klick speech.

Due to the fact that the start and end stresses of phones in a natural speech differ from comparable, but embedded phones, it is to be preferred to use corresponding audio segment bands, whose starts reproduce the start of the speech to be synthesised and whose ends reproduce the end of same, respectively.

In particular in the generation of synthesised speech, a fast and efficient procedure is desirable. For this purpose, it is to be preferred to carry out the inventive co-articulation-specific concatenation invariably by means of hard fades, with only such audio segment bands being used whose starts always reproduce a dynamic sound or phone, respectively. Such audio segment bands can be generated in advance according to the invention by means of the co-articulation-specific concatenation of corresponding audio segment bands.

In addition, the invention provides voice signals which have a natural flow of speech, speech melody, and speech rhythm, in that audio segment bands are processed before and/or after the concatenation in their entirety or in individual bands by means of suitable functions. It is particularly advantageous to perform this variation additionally in areas in which the corresponding instances of concatenation are set in order to change i.a. the frequency, duration, amplitude, or spectrum.

An still further improved signal characteristic can be achieved if the concatenation instances are set in places of the audio segment bands to be linked, where these are in agreement with respect to one or several properties.

In order to permit a simple utilisation and/or further processing of the inventive voice signals by means of known methods or devices, such as a CD player, it is to be preferred in particular that the voice signals are convertible to acoustical signals or are storable in a data carrier.

For the purpose of applying the invention also to known devices such as a personal computer or a computer-controlled musical instrument, a data carrier is provided which contains a computer program which enables the performance of the inventive method or the control of the inventive device and its various embodiments, respectively. In addition, the inventive data carrier also permits the generation of voice signals which comprise co-articulation-specific concatenations.

For providing an inventory comprising audio segments, by means of which synthesised acoustical data, in particular synthesised voice data, can be generated which does not differ from corresponding natural acoustical data, the invention provides a data storage which includes audio segments which are suited for being inventively concatenated to synthesised acoustical data. Preferably, such a data carrier includes audio segments which are suited for the performance of the inventive method, for application in the inventive device, or the inventive data carrier. Alternatively, the data carrier can also include inventive voice signals.

In addition, the invention makes it possible to provide inventive synthesised acoustical data, in particular synthesised voice data, which can be utilised with conventional devices, e.g. a tape recorder, a CD player, or a PC audio card. For this purpose, a sound carrier is provided which comprises data which at least partially has been generated by the inventive method or by means of the inventive device or by using the inventive data carrier or the inventive data storage, respectively. The sound carrier may also comprise data which are the inventively co-articulation-specific concatenated voice signals.

Further properties, characteristics, advantages, or modifications of the invention will be explained with reference to the following description; in which:

Fig. 1a is a schematic representation of an inventive device for the generation of synthesised acoustical data;
Fig. 1b shows the structure of a sound/phone;
Fig. 2a shows the structure of a conventional audio segment according to the state of the art, consisting of portions of two phones, i.e. a diphone for voice. It is essential that the solo articulation bands each are included only partially in the conventional diphone audio segment.
Fig. 2b shows the structure of an inventive audio segment which reproduces portions of a sound/phone with downstream co-articulation bands (for voice a quasi 'displaced' diphone);
Fig. 2c shows the structure of an inventive audio segment which reproduces portions of a sound/phone with upstream co-articulation bands;
Fig. 2d shows the structure of an inventive audio segment which reproduces portions of a sound/phone with downstream co-articulation bands and includes additional bands;
Fig. 2e shows the structure of an inventive audio segment which reproduces portions of a sound/phone with upstream co-articulation bands and includes additional bands;
Fig. 2f shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with downstream co-articulation bands each. The sounds/phones 2 to (n-1) each are completely included in the audio segment.
Fig. 2g shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with upstream co-articulation bands each. The sounds/phones 2 to (n-1) each are completely included in the audio segment.

Fig. 2h shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with downstream co-articulation bands each and includes additional bands. The sounds/phones 2 to (n-1) each are completely included in the audio segment.
Fig. 2i shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with downstream co-articulation bands each and includes additional bands. The sounds/phones 2 to (n-1) each are completely included in the audio segment.
Fig. 2j shows the structure of an inventive audio segment which reproduces a portion of a sound/phone of the start of a sound sequence/phone sequence;
Fig. 2k shows the structure of an inventive audio segment which reproduces portions of sounds/phones of the start of a sound sequence/phone sequence;
Fig. 2l shows the structure of an inventive audio segment which reproduces a sound/phone of the end of a sound sequence /phone sequence;
Fig. 3a shows the concatenation according to the state of the art by means of an example of two conventional audio segments. The segments begin and end with portions of the solo articulation bands (generally half of same).
Fig. 3aI shows the concatenation according to the state of the art. The solo articulation band of the middle phone comes from two different audio segments.
Fig. 3b shows the concatenation according to the inventive method by means of an example of two audio segments, each of which containing a sound/phone with downstream co-articulation bands. Both sounds/phones come from the centre of a phone unit sequence.
Fig. 3bI shows the concatenation of these audio segments by means of a cross fade.
The solo articulation band comes from an audio segment. The transition between the audio segments is effected between two

bands and is therefore less susceptible to variations (in spectrum, frequency, amplitude, etc.). The audio segments can also be processed by means of additional transfer functions prior to the concatenation.

Fig. 3bII shows the concatenation of these audio segments by means of a hard fade;

Fig. 3c shows the concatenation according to the inventive method by means of an example of two inventive audio segments, each of which containing a sound/phone with downstream co-articulation bands, with the first audio segment coming from the start of a phone sequence.

Fig. 3cI shows the concatenation of these audio segments by means of a cross fade;

Fig. 3cII shows the concatenation of these audio segments by means of a hard fade;

Fig. 3d shows the concatenation according to the inventive method by means of an example of two inventive audio segments, each of which containing a sound/phone with upstream co-articulation bands. Both audio segments come from the centre of a phone sequence.

Fig. 3dI shows the concatenation of these audio segments by means of a cross fade. The solo articulation band comes from an audio segment.

Fig. 3dII shows the concatenation of these audio segments by means of a hard fade;

Fig. 3e shows the concatenation according to the inventive method by means of an example of two inventive audio segments, each of which containing a sound/phone with downstream co-articulation bands, with the last audio segment coming from the end of a phone sequence;

Fig. 3eI shows the concatenation of these audio segments by means of a cross fade;

Fig. 3eII shows the concatenation of these audio segments by means of a hard fade;

Fig. 4 is a schematic representation of the steps of the inventive method for the generation of synthesised acoustical data.

The reference numerals used in the following refer to Fig. 1a and the numbers of the various steps of the method used in the following refer to Fig. 4.

In order to convert for example a text to synthesised speech by means of the invention, it is necessary to divide this text in a preparatory step into a sequence of phonetic characters or phonema, respectively. Preferably, prosodic information corresponding to the text is to be generated as well. The sound or phone sequence, respectively, as well as the prosodic and additional information serve as input values for the inventive method or the inventive device, respectively.

The sounds/phones to be synthesised are supplied to an input unit 101 of the device 1 for the generation of synthesised voice data and stored in a first memory unit 103 (see Fig. 1a). By means of a selection means 105 audio segments are selected from an inventory including audio segments (elements) which is stored in a database 107, or by an upstream synthesis means 108 (which is not part of the invention), which reproduce sounds or phones, respectively, or portions of sounds or phones, respectively, which correspond to the individually input phonetic characters or phonema, respectively, or portions of same and stored in a second memory unit 109 in an order corresponding to the order to the input phonetic characters or phonema, respectively. If the inventory includes portions of phone sequences or of audio segments, the selection unit 105 preferably selects those audio segments which reproduce the highest number of portions of the phone sequences or polyphones, respectively, which correspond to a sequence of phonetic characters or phonema, respectively, from the input phone sequence or phoneme sequence, respectively, so that a minimum number of audio segments is required for the synthesis of the input phoneme sequence.

If the database 107 or the upstream synthesis means 108 provides an inventory with audio segments of different types, the selection means 105 preferably selects the longest audio segment bands which reproduce portions of the sound sequence/ phone sequence in order to synthesise the input sound sequence or phone sequence, respectively, and/or a sequence of sounds/ phones from a minimum number of audio segment bands. In this context, it is advantageous to use audio segment bands reproducing linked sounds/phones, which reproduce an earlier static sound/phone and a later dynamic sound phone. In this manner, audio segments are generated which, because of the embedded dynamic sounds/phones invariably begin with a static sound/ phone. For this reason, the concatenation procedure for such audio segments is simplified and standardised, because only cross fades are required for this.

In order to achieve a co-articulation-specific concatenation of the audio segment bands to be linked, the concatenation instances of two successive audio segment bands are established with the aid of a concatenation means 111 as follows:

- If an audio segment band is to be used for synthesising the start of the input sound sequence/phone sequence (step 1), an audio segment band is to be selected from the inventory, which reproduces the start of a sound sequence/phone sequence and to be linked with a later audio segment band (see Fig. 3c and step 3 in Fig. 4).

- In the concatenation of a second audio segment band with an earlier first audio segment band, a distinction must be made as to whether the second audio segment band starts with the reproduction of a static sound/phone or a dynamic sound/phone in order to appropriately make the selection of the instance of concatenation (step 6).

- If the second audio segment band starts with a static sound/ phone, then the concatenation is carried out in the form of a cross fade, with the instance of concatenation being set in the downstream portion of the first audio segment band and in the upstream portion of the second audio segment band, with the two bands overlapping in the concatenation or at least bordering on one another (see Figs. 3bI, 3cI, 3dI, and 3eI; concatenation by means of cross fade).

- If the second audio segment band starts with a dynamic sound /phone, then the concatenation is carried out in the form of a hard fade, with the instance of concatenation being set immediately after of the downstream portion of the first audio segment band and immediately before the upstream band of the second audio segment band (see Figs. 3bII, 3cII, 3dII, and 3eII; concatenation by means of hard fade).

In this manner, new audio segments can be generated from the originally available audio segment bands, which start with the reproduction of a static sound/phone. This is achieved in that audio segment bands which start with the reproduction of a dynamic sound/phone are linked later with audio segment bands which start with the reproduction of a static sound/phone. Though this increases the number of audio segments or the volume of the inventory, respectively, can, however, be a computational advantage, because fewer individual concatenations are required for the generation of a phone sequence/ phoneme sequence, and concatenations have to carried out only in the form of cross fades. Preferably, the new linked audio segments are supplied to the database 107 or another memory unit 113.

A further advantage of this linking of the original audio segment bands to new longer audio segments results if, for example, a sequence of sounds/phones frequently repeats itself in the input sound sequence/phone sequence. It is then possible to utilise one of the new correspondingly linked audio segments, and it is not necessary to carry out another concatenation of the originally available audio segment bands with each occurrence of this sequence of sounds/phones. Preferably, overlapping co-articulation effects, too, are to be covered, or specific co-articulation effects in the form of additional data is to be assigned to the stored linked audio segment, respectively, when storing such linked audio segments.

If an audio segment band is to be used for synthesising the end of the input sound sequence/phone sequence, an audio segment band is to be selected from the inventory, which reproduces an end of a sound sequence/phone sequence, and to be linked with an earlier audio segment band (see Fig. 3e and step 8 in Fig. 4).

The individual audio segments are stored in a coded form in the database 107, with the coded form of the audio segments, apart from the waveform of the respective audio segment, being able to indicate which type of concatenation (e.g. hard fade, linear or exponential cross fade) is to be carried out with which later audio segment band, and at which instance the concatenation takes place with which later audio segment band. Preferably, the coded form of the audio segments also includes information with respect to the prosody, higher-order co-articulations and transfer functions which are used to achieve an additional improvement of the voice quality.

In the selection of the audio segment bands for synthesising the input sound sequence/phone sequence, the audio segment bands selected as the later ones are such that they correspond to the properties of the respective earlier audio segment bands, i.a. type of concatenation and concatenation instance. After the selection of the audio segment bands, each of which reproducing portions of the sound sequence/phone sequence, from the database 107 or the upstream synthesising means 108, the concatenation of two successive audio segment bands by means of the concatenation means 111 is carried out as follows. The waveform, the type of concatenation, the concatenation instance as well as any additional information, if required, of the first audio segment band and the second audio segment band are loaded from the database of the synthesising means (Fig. 3b and steps 10 and 11). Preferably such audio segment bands are selected in the above mentioned selection of the audio segment bands, which are in agreement with each other with respect to their type and instance of concatenation. In this case, loading of information with respect to type and instance of concatenation of the second audio segment band is no longer necessary.

For the concatenation of the two audio segment bands, the waveform of the first audio segment band in a later band and the waveform of the second audio segment band in an earlier band, each are processed by means of suitable transfer functions, e.g. multiplied by a suitable weighting function (see Fig. 3b, steps 12 and 13). The lengths of the later band of the first audio segment and of the earlier band of the second audio segment result from the type of concatenation and the time position of the concatenation instance, with these lengths also being able to be stored in the coded form of the audio segments in the database.

If the two audio segment bands are to be linked by means of a cross fade, they are added in an overlapping manner according to the respective instance of concatenation (see Figs. 3bI, 3cI, 3dI, and 3eI; step 15). Preferably, a linear symmetrical cross fade is to be used herein, however, any other type of cross fade or any type of transfer function can be employed as well. If a concatenation in the form of a hard fade is to be carried out, the two audio segment bands are not joined consecutively in an overlapping manner (see Figs. 3bII, 3cII, 3dII, and 3eII; step 15). As can be seen in Fig. 3bII, the two audio segment bands are arranged immediately successive in time. In order to be able to further process the voice generated in this manner, it is preferably stored in a third memory unit 115.

For the further linking with successive audio segment bands, the audio segments bands linked so far are considered as a first audio segment band (step 16), and the above described linking process is repeated until the entire sound sequence/ phone sequence has been synthesised.

For an improvement of the quality of the synthesised voice data, the prosodic and additional information which are input in addition to the sound sequence/phone sequence, are preferably to be considered in the linking of the audio segment bands. By means of known methods, the frequency, duration, amplitude, and/or spectral properties of the audio segment bands can be modified before and/or after the concatenation in such a manner that the synthesised voice data comprises a natural word and/or phrase melody (steps 14, 17, or 18). In this context it is to be preferred to select concatenation instances at places of the audio segment bands, at which they agree in one or several suitable properties.

In order to optimise the transitions between two successive audio segment bands, the processing of the two audio segment bands by means of suitable functions in the area of the concatenation instance is additionally provided, in order to i.a. tune the frequencies, durations, amplitudes, and spectral properties. The invention additionally permits to take into consideration higher-order acoustical phenomena of a real speech, such as for example higher-order co-articulation effects of style of speech (i.a. whispering, stress, singing voice, falsetto, emotional expression) in the synthesising of the sound sequence/phone sequence. For this purpose, information relating to such higher-order phenomena, is additionally stored in a coded form with the corresponding audio segment bands in order to select only such audio segment bands in the selection which correspond to the higher-order co-articulation properties of the earlier and/or later audio segment bands.

The synthesised voice data generated in this manner preferably have a form which, with the aid of an output means 117, allows to convert the voice data to acoustical voice signals and to store the voice data and/or voice signals in an acoustical, optical, magnetic, or electrical data carrier (step 19).

Generally, inventory elements are generated via the recording of actually spoken speech. Depending on the level of training of the inventory-building speaker, i.e. his or her capability for controlling the speech to be recorded (e.g. to control the tone level of the speech or to speak exactly on one tone level), it is possible to generate identical or similar inventory elements which have displaced boundaries between the solo articulation bands and the co-articulation bands. This results in considerably more possibilities of setting the concatenation points in different places. As a consequence, the quality of a speech to be synthesised can be considerably enhanced.

This invention allows for the first time to generate synthesised voice signals by means of a co-articulation-specific concatenation of individual audio segment bands, because the instance of concatenation is selected depending on the respective audio segment bands to be linked. In this manner, a synthesised speech can be generated which is no longer distinguishable from a naturally spoken speech. Contrary to known methods or devices, the audio segments used herein are not generated by speaking or recording, respectively, complete words, in order to ensure an authentic voice quality. It is therefore possible by means of this invention to generate synthesised speech of any contents with the quality of an actually spoken speech.

Although this invention is described by way of the example of the speech synthesis, it is not limited to the field of synthesised speech, but can be used for synthesising any acoustical data or any sound events, respectively. This invention can therefore be employed for the generation and/or provision of synthesised voice data and/or voice signals for any language or dialect, as well as for the synthesis of music.